NORC at the University of Chicago

FINAL REPORT
January 2026

PPRL2-23-N02:
Final Report

Presented by:
NORC at the University of Chicago

Presented to:
National Center for Science and Engineering Statistics within the U.S. National Science Foundation

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-014).

Table of Contents

Executive Summary

The National Secure Data Service Demonstration (NSDS-D) is a federal initiative authorized under Section 10375 of the CHIPS and Science Act of 2022. The initiative is operated by the National Center for Science and Engineering Statistics (NCSES), which is a federal statistical agency in the U.S. National Science Foundation (NSF), legislatively mandated in the National Science Foundation Act of 1950 (42 U.S.C. 1862 (a) (6)) to serve as a central federal clearinghouse for the collection, interpretation, analysis, and dissemination of objective data on science, engineering, technology, and research and development and to provide a source of information for policy formulation by other agencies of the federal government.

The NSDS-D is designed to inform the development of a future government-wide shared service for data linkage, data sharing, secure access, and user training. This shared-services model aims to enhance evidence-based policymaking across the federal government while maintaining strong protections for privacy and confidentiality. Many NSDS-D activities such as this project are sponsored through America's Data Hub Consortium (ADC), supporting cross-sector collaboration in areas such as data linkage, secure access, privacy, and data infrastructure.

This report documents one such demonstration project with a primary goal to inform an understanding of the infrastructure needed to implement privacy enhancing technology to link data without sharing unencoded personally identifiable information (PII) while addressing appropriate data protection and governance. This project serves as a case study to explore the feasibility of linking data from a federal statistical agency (NCSES) with its parent agency (NSF) using Privacy-Preserving Record Linkage (PPRL) techniques, specifically by connecting NCSES's Survey of Earned Doctorates (SED) with NSF's Principal Investigator (PI) award data. NORC at the University of Chicago was contracted to perform the linkage work and served as the trusted third party throughout the project.

To achieve this, a Data Sharing Agreement (DSA) was established between NCSES and NSF's Office of the Chief Information Officer (OCIO) where NORC was identified as the trusted third party to perform the linkage within the NCSES Secure Data Access Facility (SDAF). The project evaluated two PPRL tools, Datavant and Anonlink, ultimately selecting Anonlink. The linkage process involved preprocessing and encoding data within each entity's secure environment, transferring only encoded identifiers to the NCSES SDAF, and performing the linkage using Anonlink. The output consisted of a linkage status file and a linked analysis dataset. This linked analysis dataset enabled new statistical analyses that were previously unfeasible, such as tracking the time from doctoral graduation to the first NSF award and examining award frequency across various demographic and field of study subgroups.

The report concludes with recommendations to guide future PPRL initiatives in a shared service environment. It underscores the significance of adopting standardized linkage frameworks, engaging multidisciplinary teams, and establishing explicit data sharing agreements. It also highlights the need for robust preprocessing safeguards and secure environments that align with the data encoding methods implemented. Overall, the project demonstrates both the technical feasibility and the strategic value of PPRL within the federal statistical system, offering a replicable model for future cross-agency collaborations.

Introduction

Background and Purpose

This project was initiated by the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF) as part of the National Secure Data Service (NSDS) Demonstration effort. Its core motivation was to evaluate services that can be offered through a shared service environment to increase access to data that otherwise may not be possible without these tools. Specifically, the project aimed to demonstrate the feasibility of linking data from a federal statistical agency (NCSES) with its parent agency (NSF) [1] using Privacy-Preserving Record Linkage (PPRL) techniques. The linkage focused on combining the NCSES Survey of Earned Doctorates (SED) with NSF Office of the Chief Information Officer (OCIO) Principal Investigator (PI) award data. This effort helps enable secure, cross-agency data integration without exposing personally identifiable information (PII), thereby supporting evidence-based policymaking and program evaluation.

As an NSDS Demonstration Project, this project serves as a testbed for developing replicable methods and infrastructure that could be scaled across the federal statistical system. The project established a Data Sharing Agreement (DSA) between NCSES and NSF OCIO, selected Anonlink as the PPRL software, and designated NORC as the trusted third party to perform the linkage within the NCSES Secure Data Access Facility (SDAF). The demonstration not only validated the technical feasibility of PPRL but also produced analysis datasets that enable new statistical insights. The lessons learned from this project are expected to inform a future NSDS, particularly in establishing the infrastructure to implement secure and effective linkage processes, refine disclosure risk mitigation strategies, and create new linked data assets to be used for research and policy development.

Report Outline

This report is organized to reflect the full lifecycle of the SED-PI linkage demonstration project, from data sharing agreements to technical implementation and analysis outcomes. It begins with the development of a DSA between NCSES and NSF OCIO, establishing the legal and operational framework for secure data collaboration. Subsequent sections detail the evaluation and selection of PPRL software, the preprocessing and encoding strategies employed to protect PII, and the linkage methodology implemented using Anonlink. The report then presents high level linkage results and statistical analyses conducted using the linked SED-PI dataset, followed by recommendations of how to offer PPRL services to guide future NSDS efforts. Each section is designed to provide transparency, technical rigor, and actionable insights for replicability across federal data linkage initiatives.

[1] "The term "parent agency" means every organizational level of an agency, including sub-agencies, offices, components, or units, as well as any organizational units that contain a Recognized Statistical Agency or Unit, but the term does not include the Recognized Statistical Agency or Unit itself" (OMB, 2023).

Developing a Data Sharing Agreement (DSA)

DSA Development

A DSA was created to define the terms of collaboration between NCSES, the statistical agency, and NSF OCIO, the parent agency. This agreement focused on developing methods and infrastructure to facilitate the SED-PI data linkage.

Developed collaboratively by representatives from the different entities, the DSA emerged through iterative consultations to ensure alignment with federal data governance standards and privacy-preserving principles. Key considerations included the legal requirements of NCSES and NSF to safeguard the PII of each entity's data sources. Although both entities reside within the same parent organization, the DSA formalized roles, responsibilities, and data handling protocols to uphold the integrity of the linkage process. This structure serves as a model for other projects involving parent-statistical agency relationships, highlighting the necessity of formal agreements even within a shared institutional framework to reinforce transparency, accountability, and trust in intra-agency collaborations. The DSA was developed as a modular resource that can be used to guide future efforts to negotiate data sharing agreements across the federal statistical system.

DSA Content

The DSA outlines the purpose, authorities, and background, along with the responsibilities of each party, the data linkage and access workflow, and other relevant provisions. Below is an overview of the key content.

To facilitate the testing and evaluation of the PPRL technology, the DSA outlines the responsibilities of each party involved. Exhibit 1 presents a table detailing these responsibilities at each stage of the linkage process. Through the process we learned the importance of having visual prototypes of the process as different reviewers wanted to see information displayed differently. Exhibit 2 illustrates the key steps in the PPRL process, specifying when and where these should be carried out. Specifically, the DSA notes that NCSES and NSF OCIO are responsible for encoding the SED and PI data, respectively, and will only transfer PII that has been encoded to the NCSES SDAF. Within the NCSES SDAF, NORC, as the trusted third-party, is responsible for performing the linkage using the encoded data and creating the linked SED-PI dataset.

Additionally, the DSA details the confidentiality and data security provisions that protect the SED and PI data, including the Privacy Act of 1974 and the NSF Act of 1950 and other policies that may impact data sharing. As a result, these provisions also apply to the linked SED-PI dataset.

Exhibit 1. DSA Responsibilities

	NCSES SED PII and Analytic data files	**NSF PI PII and Analytic data files**	**NCSES SED Encrypted PII/Analytic data files**	**NSF PI Encrypted PII/ Analysis data files**	**Linked Analysis dataset**
	NCSES secure environment	NSF secure environment	NCSES SDAF	NCSES SDAF	NCSES SDAF Project
NCSES programmer*	X				
NSF OCIO programmer**		X			
Trusted third party contractor for NCSES – NSF PPRL project (NORC)			X	X	X
Approved researchers***					X

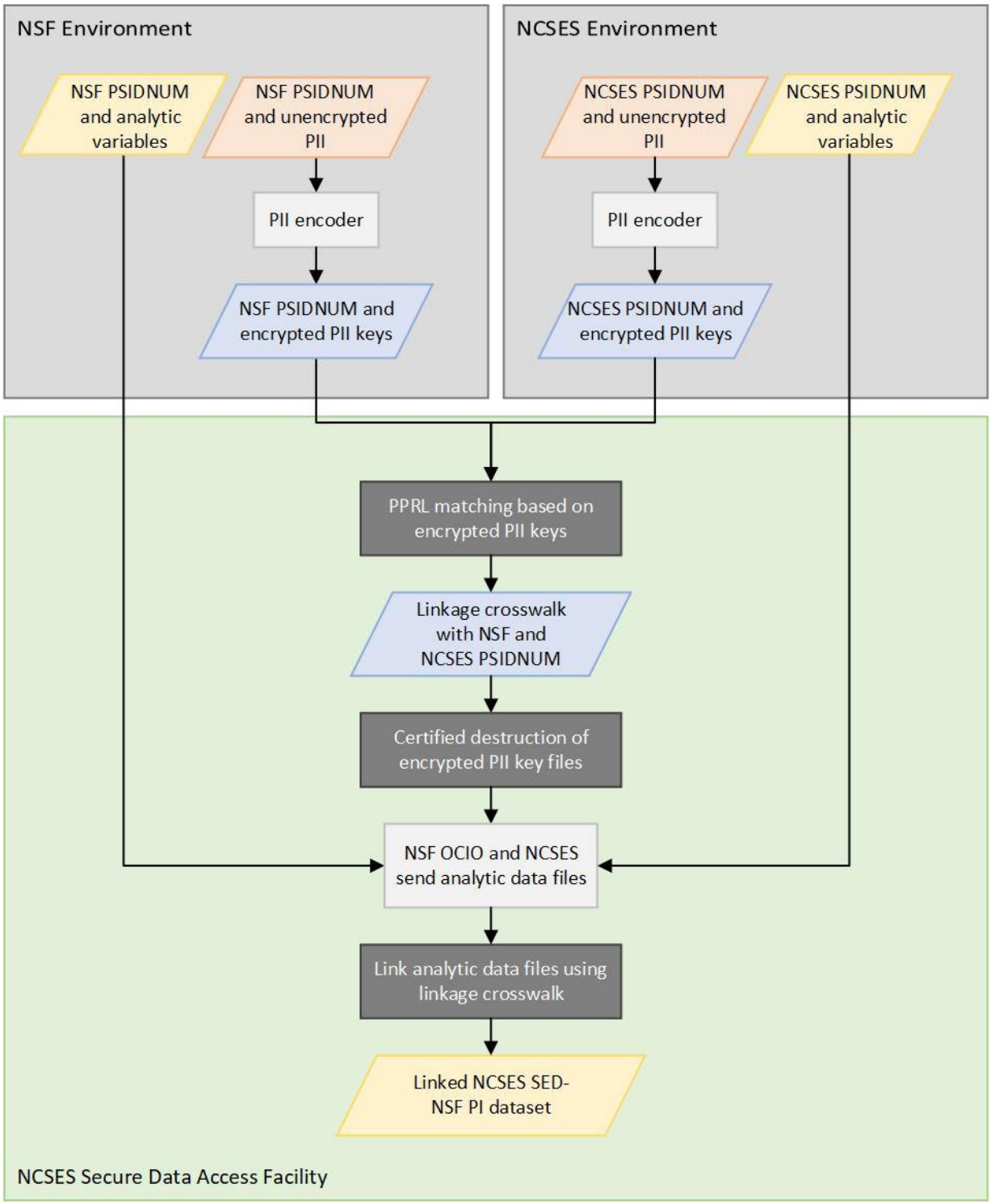
*NCSES employee or an onboarded contractor or pathways fellow

**NSF OCIO to determine programmer role

***may include NSF and/or NCSES researchers

Definitions: NCSES is the National Center for Science and Engineering Statistics. SED is the Survey of Earned Doctorates. NSF is the National Science Foundation. OCIO is the Office of the Chief Information Officer. PI is Principal Investigator. PII is Personally Identifiable Information. SDAF is the NCSES Secure Data Access Facility.

Exhibit 2. Data Linkage Process Workflow



Note: Multiple encrypted keys may be generated to utilize different combinations of PII data.

Definitions: NCSES is the National Center for Science and Engineering Statistics. SED is the Survey of Earned Doctorates. NSF is the National Science Foundation. OCIO is the Office of the Chief Information Officer. PI is Principal Investigator. PII is Personally Identifiable Information. PSIDNUM is a project specific identification number. PPRL is privacy-preserving record linkage.

Software Selection

Selecting a suitable tool is crucial for a successful PPRL effort, as it ensures the implementation is efficient, accurate, and compliant with data security needs. There are a variety of PPRL tools available, including open-source and commercial software. This project evaluated two widely used tools, Datavant and Anonlink. An internal Software Selection Recommendations Report (NORC, 2024) was developed presenting multiple factors relevant to the software selection process and providing a recommendation for the PPRL tool best suited to meet the requirements of the SED-PI linkage.

Evaluated Tools

Two PPRL tools were evaluated: Datavant and Anonlink. Datavant is a commercially licensed PPRL tool created primarily for linking health-related data sources (Datavant, 2023-b). Datavant is a FedRAMP authorized solution (Datavant, 2023-a) that encodes records into encrypted tokens to facilitate a secure PPRL. Anonlink is an open-source Python software (optimized in C++) under Apache 2.0 developed by CSIRO's Data61 and is available on GitHub (CSIRO, 2017). Anonlink implements PPRL encoding records into cryptographic longterm keys (CLKs) as described by Schnell, Bachteler, and Reihner (2011). Anonlink's CLKs are Bloom filters[2] created with the hashing functions SHA-256 or SHA-512, approved as secure hashing algorithms by the National Institute of Standards and Technology (NIST, 2015). Both Datavant and Anonlink have been successfully used by other federal agencies to conduct PPRL projects (ASPE, 2024; Mirel et al., 2022).

Tool Selection Considerations

A range of considerations were identified to evaluate the two PPRL tools and select the most appropriate option.

First, the availability of variables that could be used for linkage across datasets was considered. An analysis of shared fields between the SED and PI data revealed limited but sufficient overlap to support linkage; however, the absence of unique identifiers—such as date of birth—limited high-precision linkage determination.[3] The available data were insufficient to meet the requirements for any of the standard Datavant core tokens as defined (Datavant, 2022). This would have necessitated adjustments to these tokens or the creation of custom tokens. However, without extensive analysis (as Datavant has already done with its core tokens) it would not be possible to determine the relative or combined reliability of token agreements.

[2] Bloom filter encoding is a widely used technique for linking sensitive databases, first introduced in the context of PPRL by Schnell et al. (2011).

[3] The following variables were available both in SED and PI data sources and were considered as potential linkage identifiers: First Name, Last Name, Middle Name Initial, Name Suffix, Last 4 Digits of Social Security Number, Ph.D. Graduation Year, Sex, Ethnicity, U.S. Citizenship.

Second, the location of PPRL activities was considered. Establishing secure environments for PPRL is essential to protect sensitive data and comply with privacy regulations. NSF may require an Authority to Operate (ATO) before transferring restricted data to non-NSF environments, including FedRAMP-compliant ones—a process that can be time-consuming. When evaluating the tools, Datavant allows local encryption within the client's environment but may still control the encryption seed and may require external environments to perform the linkage, potentially triggering ATO requirements. Anonlink's open-source tools can be fully operated within NSF OCIO and NCSES secure environments, with NSF OCIO and NCSES programmers performing the encoding before transferring the CLKs to the NCSES SDAF for linkage.

Furthermore, disclosure risk was assessed. Datavant offers Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (HHS, 2025)[4] compliant expert determinations for its standard encrypted tokens, but using modified or custom encrypted tokens could invalidate these determinations. Although Anonlink is open-source, enabling transparent code review, and it utilizes NIST-approved hashing algorithms, it does not include HIPAA Privacy Rule expert determinations for the encrypted CLKs. Therefore, additional disclosure protection steps were added to help ensure compliance with HIPAA privacy requirements (e.g., separating covariates from CLKs) and maintaining all linkage activities within a secure compute environment.

Taken together, these considerations led to the selection of Anonlink for this case study. Most importantly, the project's data limitations—particularly the absence of date of birth and other strong unique identifiers—limited the use of Datavant's standard, validated token framework without resorting to unvalidated custom tokens. In contrast, Anonlink's methodology allows CLKs to be constructed from the fields that are available, without being constrained by predefined token specifications, while also allowing all processing to remain within NSF and NCSES secure environments.

Final Tool Recommendation

Given the considerations summarized above, NORC recommended the use of Anonlink as the PPRL software for the SED-PI linkage. It is important to note, however, that this choice was specific to the needs and context of this project; the selection of PPRL software should always be guided by the characteristics of the data and the requirements of each linkage project.

Following the decision to use Anonlink, NORC was established as the trusted third party, responsible for completing the linkage within the NCSES SDAF. NCSES and NSF OCIO were responsible for encoding the data, ensuring that PII was only transferred to the NCSES SDAF after encoding, thereby preventing NORC from accessing the clear-text PII.

[4] While the HIPAA Privacy Rule specifically governs the protection of personal health information, HIPAA is referenced here because its risk-assessment principles and de-identification standards are widely recognized and can serve as a useful conceptual benchmark. Although this linkage project does not involve protected health information, similar evaluative frameworks can be applied to non-health datasets to promote rigorous privacy and disclosure-risk assessment.

Summary of Disclosure Risk Mitigation Strategies

Additional disclosure risk mitigation strategies specific to the use of Anonlink for the SED-PI linkage were implemented. A thorough security review of the Anonlink Python package was conducted by the NORC Data Enclave (NORC, 2025-c) security team before installation, assessing software vulnerabilities, compatibility, and developer reputation. Data custody and server security protocols were designed to ensure that no clear-text PII left the NSF OCIO or NCSES environments. NORC provided guidance on creating CLKs without accessing clear-text PII, and all data transfers to the NCSES SDAF were securely conducted. Administrative safeguards, including signed DUAs and CIPSEA agent training, were also enforced.

To address the risk of intruder attacks on CLKs, mitigation strategies included using composite CLKs with multiple fields, applying confidential salt[5] and secret[6] values for hashing that were only accessible to project team members behind their file walls and not shared with the trusted third party, and restricting user access to the linkage environment within the NCSES SDAF only to project team members. The linkage process involved encoding, secure transmission, and deletion/destruction of CLKs before transferring covariate data for analysis. The final linkage files, while free of direct identifiers, contained covariate data with indirect identifiers, prompting a disclosure review process for any statistical products derived from the data exported from the NCSES SDAF. Access to the linked SED-PI dataset will be restricted and governed by a formal application and licensing process as outlined in the DSA. These comprehensive measures provided strong safeguards against disclosure risk throughout the project, even in the absence of a HIPAA Privacy Rule expert determination for encoded CLKs.

Linkage Methodology

This project included the development of an internal Linkage Approach Report (NORC, 2025-a) prior to linkage execution that defined the methodology in detail. This internal report established the procedural framework—including the steps summarized in this section—and served as the foundation for implementing the linkage in Python using the Anonlink library.

Data and Preprocessing

This demonstration project leverages two datasets:

[5] A salt is a randomly generated value added to input data before hashing, used to ensure that identical inputs produce different hashes and to thwart precomputed attacks (Rosulek, 2021).

[6] In Anonlink, the secret is a shared cryptographic key known only to the data parties, used to generate the CLKs from the PII and must remain confidential (CSIRO's Data61, 2020).

- **NCSES Survey of Earned Doctorates (SED)**: This dataset is an annual census of individuals who received research doctorates from accredited U.S. institutions. For the purposes of this demonstration, the SED dataset includes only records from doctorate recipients between 2012 and 2022. These records contain linkage identifiers such as name, SSN-4 (last four digits of the Social Security Number), and Ph.D. graduation year, as well as covariates related to demographics, educational background, and Ph.D. field of study.

- **NSF OCIO Principal Investigator (PI) Award Data**: This dataset includes administrative information on PIs that received NSF research awards. For this project, only PIs with awards received between 2012 and 2022 were considered. The PI data include linkage identifiers such as names, SSN, and degree year, which align with the SED dataset and enable linkage. It also contains covariates capturing award details, demographics, and institution characteristics.

To guide the selection of linkage identifiers and preprocessing steps, NCSES and NSF OCIO conducted a data quality assessment with guidance from NORC. This included quality control (QC) checks to evaluate the completeness, consistency, and validity of the identifiers. Based on these findings, the following variables were selected as linkage identifiers: first name, middle name initial, last name, SSN-4, and degree year. Sex and ethnicity were used as checks with linkage accuracy. Race was evaluated as a potential variable for assessing linkage accuracy; however, there was substantial missing data for this field. As a result, race was not used for the linkage assessment. The data quality checks also identified preprocessing needs—such as name data standardization—to ensure comparability across data sources.

Based on the results of this assessment, NORC developed preprocessing code and guidance for NCSES and NSF OCIO to perform on the data. Data preprocessing involved deduplicating PI awards data to the person level, adding project-specific IDs assigned to protect internal identifiers, cleaning and standardizing fields, and excluding records with missing key identifiers. Additionally, a set of synthetically generated test records were introduced to confirm consistent encoding since this was the first time Anonlink was used with these data sources.

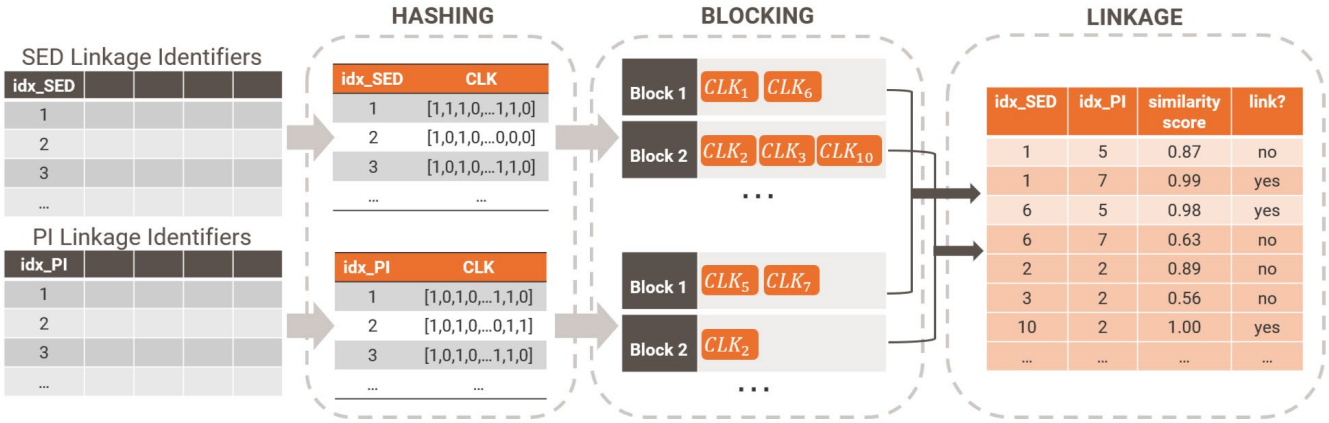
Finally, alternate records were generated for multi-part last names to account for possible variations in name representation across datasets[7]. For SED, additional alternates were created based on first and middle name. In total, 9 percent of SED records and 3 percent of PI records included alternate versions. Based on the final linkage results, the creation of alternate records improved the completeness of linkage by over 5 percent and therefore was considered useful to the linkage process.

[7] For example, consider the fictitious individual ‘Carlos Ramirez Lopez’, where ‘Ramirez Lopez’ is the last name. To account for variations in how this two-part last name might appear across datasets, three records would be generated — one with Ramirez, one with Lopez, and one with Ramirez Lopez.

Linkage Workflow

Exhibit 3 illustrates the Anonlink PPRL workflow applied in this project, detailing the transformation of clear-text data into CLKs to protect confidentiality, the implementation of a proposed blocking strategy to optimize efficiency, and the linkage procedure in which CLKs were compared to establish links and generate the linked SED-PI dataset.

Exhibit 3. PPRL Methodology Workflow



Definitions: SED is the Survey of Earned Doctorates. PI is Principal Investigator. CLKs are cryptographic long-term keys corresponding to the encoded data. Similarity score is the Sørensen–Dice coefficient used to measure similarity between two CLKs.

Data Encoding

Anonlink utilizes a privacy-preserving method by converting linkage identifiers into CLKs using Bloom filter encoding. This technique, originally introduced by Schnell et al. (2011), hashes elements from data fields into a binary bit vector via multiple independent hash functions. Two versions of CLKs were generated—one including the SSN-4 field and one excluding it—due to SSN-4's partial missingness in the data to improve linkage accuracy. Each version employed a distinct hashing schema with separate salt (Rosulek, 2021) and secret (CSIRO's Data61, 2020) keys to enhance security, and specifications on hashing configurations tailored for each data field, e.g., 2-gram comparison for name fields and exact comparison for SSN-4.

The hashing schema also specified the number of bits assigned per linkage identifier, which was determined using weight differentials derived from Fellegi-Sunter agreement and disagreement weights, based on m- and u-probabilities. These probabilities were sourced from published values, where available. NORC provided an encoding code package that converted the data into CLKs using Anonlink software for NCSES and NSF OCIO to apply to the SED and PI data, respectively. The final submission files consisted of CLKs and their assigned project-specific ID, with no clear-text PII. They were securely transferred to the NCSES SDAF for QC checks before proceeding with the linkage process.

Blocking

Where feasible, conducting a full record linkage without blocking allows for exhaustive comparisons between all records in the two datasets, thereby maximizing linkage accuracy. For this project, the linkage was successfully performed without blocking. However, due to the computational intensity of such an approach, blocking may be required in other PPRL efforts with larger datasets or other computational environments to reduce the number of comparisons while preserving linkage quality. Because this project was designed to serve as a testbed for developing replicable PPRL methods and infrastructure that could be scaled across the federal statistical system, NORC evaluated the feasibility of implementing two different blocking strategies for future use within the Anonlink tool.

The primary blocking method implemented was the locality sensitive hashing (LSH) λ-fold technique, implemented within the Anonlink software, which groups similar CLKs into blocks using multiple composite hash functions (Karapiperis and Verykios, 2015). Additionally, NORC tested an exploratory "blocking by bits" strategy, which created blocks based on sequential sets of CLK bits (Resnick and Adell Raventós, 2024). Note that both methods operate directly on the encoded data and could be applied within the NCSES SDAF. This ensured that no clear-text data was exposed during the blocking process. Initial testing showed that it was possible to reduce computational processing requirements without appreciable loss of linkage accuracy, particularly using the LSH λ-fold technique. Given the ability of the computing environment to meet the processing demands of the SED-PI linkage, blocking techniques were not needed and therefore were not used.

Linkage

Anonlink Output Pairs

The linkage process involved comparing candidate record pairs using Anonlink. This comparison generated a similarity score for each pair of CLKs, which corresponded to the Sørensen–Dice coefficient (Brown et al., 2019). A baseline threshold was set for this process to manage memory resources, resulting in a list of candidate pairs with similarity scores above the baseline. The output from Anonlink was initially assessed with synthetically generated test data to identify any potential encoding issues and confirm that the results were as expected.

Additionally, the Anonlink output from both sets of CLKs (with and without SSN-4) on the true data was compared as part of the QC procedures. This review identified a potential issue with the PI SSN-4 data. NCSES collaborated with NSF OCIO to investigate the PI SSN-4 data, attempting to mitigate the issue and recreate the with-SSN-4 CLKs. Due to multiple unsuccessful attempts to identify the issue, NCSES and NORC agreed to proceed with the linkage using only the without-SSN-4 CLKs. Risk mitigation strategies as part of the QC process were implemented as noted above in section "*Summary of Disclosure Risk Mitigation Strategies*."

Next, the Anonlink output candidate pairs were deduplicated, retaining only the pair with the highest similarity score for each record, so that each SED and PI record pair appeared at most once in the resulting set.

Linkage Assessment and Cutoff Threshold Selection

The final step of the linkage process involved selecting a similarity score cutoff threshold to distinguish links from non-links. To achieve this, match probability—defined as the estimated likelihood that a pair is a true match—was estimated at each similarity score level using sex agreement rates across SED and PI. Precision was also estimated at each similarity score level as the average match probability calculated for all pairs whose similarity scores exceeded the given similarity score level. Additionally, recall was estimated at each similarity score level based on SED participants that linked to PI award recipients where the awards were received within 2012-2022 and the PI reported a degree graduation year within the SED year range (2012-2022). This recall computation is based on the expectation that PI award recipients with a degree graduation year between 2012 to 2022 should have a matching record in the SED file (covering the same years) used for linkage, if the degree was from a US accredited institution[8].

Precision and recall were then combined into an F-score with $\beta = \frac{1}{2}$. This beta gives precision twice the weight as recall, which reflects the common preference in linked-data applications to prioritize linkage accuracy over completeness. The selected similarity score cutoff threshold was the one that maximized the F-score, which was subsequently used to identify the links (see Christen et al., 2023, for a comprehensive review of the F-score and its properties). The results of the linkage assessment are summarized in the Linkage Results section.

Final Linkage Files

After producing a list of links, three linkage output files were generated:

- **Anonlink Linkage Output**: Contained deduplicated SED–PI pairs with similarity scores above the baseline threshold.

[8] SED source data does not include records for doctorate recipients who received their degree from a non-accredited US institution or an international institution. Unlike degree graduation year, the institution where the PI received their degree was not available in the PI source data.

- **Linkage Status File**: Provided detailed information on linkage outcomes, linkage status for all SED records and similarity scores for linked pairs.
- **Linked Analytic File**: Included linked pairs along with relevant SED and PI covariates. It also contained SED covariates for non-linked records, enabling researchers to perform analyses on all SED participants.

Linkage Results

To evaluate the pairs generated by Anonlink, agreement rates were first calculated for common covariates—specifically sex and ethnicity—across varying similarity score levels. Agreement rates represent the percentage of record pairs where a covariate value is consistent between the SED and PI datasets (e.g., both records indicate sex = female). As expected, agreement rates for both sex and ethnicity declined as the similarity score threshold was lowered. Sex agreement served as the basis for estimating match probability, which was then used to derive precision. The optimal similarity score cutoff threshold, selected to maximize the F-score, was 83 percent. At this threshold, the estimated precision was 92.5 percent, recall was 91.5 percent, and the F-score was 92.3 percent.

Based on the final linked SED-PI dataset, 2.7 percent of SED records (those who graduated in 2012-2022) linked to a PI award recipient for PIs receiving awards between 2012-2022. Furthermore, among SED participants (those who graduated in 2012-2022), 2.2 percent linked to a PI award recipient for PIs receiving awards between 2012-2022 and whose reported degree graduation year fell within the SED data doctorate year range (2012–2022)[9]. Linkage rates were also analyzed across covariate subgroups for SED participants that were linked to a PI award between 2012-2022. Rates by SED Ph.D. graduation year revealed higher linkage rates for earlier years, likely due to increased exposure time for researchers to obtain NSF awards post-graduation.

Finally, the estimated linkage metrics (linkage rate, precision, and recall) were compared to results from a similar linkage between the Survey of Doctorate Recipients (SDR) and public PI awards data (NCSES, 2022). The SED-PI linkage rate was 2.7 percent while the SDR-PI linkage was 9 percent. The SED-PI linkage yielded estimated precision and recall metrics that exceed those observed in the earlier SDR-PI linkage effort. Specifically, the SED-PI linkage achieved an estimated precision of 92.5 percent and recall of 91.5 percent, compared to a 79 percent precision and a 59 percent recall reported for the SDR-PI linkage. Such comparisons are valuable when available, as they can serve as a reference point and help validate the linkage process. However, several caveats must be considered when interpreting these comparisons.

[9] The PI award source data contained records for all awards received between 2012 to 2022. This includes PIs whose reported graduation year is within the 2012 to 2022 timeframe as well as PIs whose reported graduation year is outside the 2012 to 2022 timeframe. Of all PIs receiving awards within 2012-2022, 77.8 percent had a reported degree graduation year outside the 2012 to 2022 timeframe. Additionally, the PI source data included awards for PIs who received their degree from non-accredited US institutions or international institutions. Unlike degree graduation year, the institution where the PIs degree was received was not available in the PI source data.

First, while the SED and SDR datasets share structural similarities, they differ substantially in population coverage: the SED represents a yearly census of doctorate recipients, whereas the SDR comprises a longitudinal sample cohort. Additionally, the timeframe of the PI data differed across both linkages, with the SED-PI linkage limited to PI awards received from 2012 to 2022 while the SDR-PI linkage included all PI awards received through spring 2020. These differences in scope and temporal alignment complicate direct comparisons of linkage performance. Second, neither linkage effort had access to a gold standard, limiting the ability to validate matches definitively. In addition, recall for the SED-PI linkage was calculated based on SED participants that linked to PI award recipients where the awards were received within 2012-2022 and the PI reported a degree graduation year within the SED year range (2012-2022). In contrast, the SDR-PI linkage metrics were calculated using survey responses, which are subject to measurement error. Taken together, while the SED-PI linkage shows stronger performance on estimated metrics, differences in methodology, data sources, and population characteristics mean that these results should be interpreted with caution.

Analyses

Once the linkage was finalized, statistical analyses using the linked SED-PI dataset were performed to demonstrate possible statistical uses, which were provided in the project's internal Statistical Analysis Report (NORC, 2025-b). The research questions analyzed included:

(1) What are the number and proportion of SED doctorate recipients who received one or more PI awards? What are the demographic characteristics (at the time of graduation) of people who are more and less likely to have received one or more PI awards?

(2) Among SED doctorate recipients who received one or more PI awards, what is the typical time to get the first PI award? How does this vary by SED doctorate recipient's characteristics (at the time of graduation)?

(3) What is the average number of awards received among researchers who received at least one award? How does the number of awards vary by person characteristics (at the time of graduation)?

The linkage of SED and PI award data enables analyses that are not feasible using either dataset independently. While the SED offers detailed educational and demographic information, and the PI award data captures research funding outcomes, their integration allows for longitudinal tracking of individuals from doctoral completion into the NSF funding ecosystem. This facilitates measurement of time-to-award since graduation and award frequency for SED doctorate recipients across various characteristics, such as degree fields and age cohorts. The linked data also supports program evaluation and policy development by enabling more comprehensive analyses of research participation patterns and outcomes.

Recommendations

Drawing from the lessons learned from this NSDS demonstration project, this section presents key insights from the SED-PI linkage, offering guidance for future federal data linkage efforts. The recommendations are grouped into thematic clusters to highlight recurring priorities, and they aim to enhance transparency, technical rigor, and operational feasibility across future initiatives aiming to link data without sharing direct identifiers.

Data Sharing Agreements

Clarify roles, responsibilities, and timelines – Ensure transparency by clearly delineating roles and expectations for all parties involved in the DSA to mitigate risk and improve accountability.

Include explicit quality-assurance mechanisms – Specify metrics and methods for assessing the quality of linked records directly within DSAs to ensure proper linkage validation and to standardize PPRL efforts.

Tool Selection and Open-Source Software

Align PPRL tool selection with context – PPRL tools are not one-size-fits-all; select solutions that align with the quality and availability of the data and the team's technical skills, as well as the data security and computation environment requirements.

Plan for customization and limited documentation when using open-source solutions – Open-source tools like Anonlink may have limited documentation and no technical support, requiring teams to troubleshoot issues and develop their own methods for settings such as bits per linkage identifier and cutoff thresholds. While this flexibility enables customization, it also demands extra time and expertise.

Mitigate risks from encoding vulnerabilities – Bloom filter-based CLKs used in Anonlink preserve similarity for probabilistic record linkage, but this also makes them susceptible to frequency and dictionary intruder attacks. In this project, risks were reduced by conducting linkage in the NCSES SDAF environment, using secret keys, and creating composite CLKs from multiple fields. Future linkages should use secure, ideally FedRAMP-compliant environments when employing Bloom filter encoding.

Data Preparation and Linkage

Strengthen pre-linkage quality safeguards – Standardize data quality checks across datasets and embed robust validation steps into workflows before encoding. For variables with QC limits (e.g., SSN-4), use alternatives such as producing multiple sets of CLKs for linkage validation and improved accuracy. Preprocessing and linkage code can be initially tested with synthetic datasets and then tested on subsets of real encoded data to catch errors before full linkage execution.

Foster clear communication and collaborative troubleshooting – For open-source workflows, improve efficiency by ensuring teams understand code and data interoperability before implementation. It is recommended to employ approaches such as tutorial meetings or virtual screen sharing, where suitable, to address coding challenges and facilitate knowledge transfer. Additionally, utilizing synthetic data can help with troubleshooting code effectively.

Standardized Frameworks for Data Linkages

Establish standardized linkage frameworks for agencies – Create formal frameworks to guide and standardize government data linkage projects. These should include QC checkpoints at each stage of the linkage to prevent data and process issues, drawing on lessons from demonstration projects such as this one to share effective, consistent protocols.

Make technical information accessible – Complex processes should be communicated in ways that non-technical audiences can understand. Using clear, high-level visuals –such as flowcharts—helps make technical documentation more accessible.

Team Composition

Assemble multidisciplinary teams – Combine expertise in record linkage, privacy protection, encryption, data sharing, project management, communications, and federal data systems to boost project efficiency and quality. It is crucial to understand the roles and responsibilities of each party and ensure they have the necessary data knowledge, technical skills, and availability. Each data source in a PPRL workflow based on a trusted third-party model should have a dedicated champion with deep data understanding, and technical skills and availability to handle data preprocessing and encoding tasks.

Conclusion

This demonstration project provides compelling evidence of the feasibility and value of PPRL using Anonlink when clear text matching is not an option. By successfully linking SED and PI award records, the project demonstrates how statistical agencies and their parent agencies—here, NCSES and NSF—can collaborate effectively while maintaining the independence and adhering to the confidentiality mandates of each entity. The DSA developed for this effort serves as a model for future initiatives, highlighting the importance of clearly defined roles, secure environments, and encoded data workflows. The linkage methodology, which included data preprocessing, Bloom filter encoding, and probabilistic record linkage via Anonlink, enabled the creation of a linked SED-PI dataset while protecting sensitive data. This linked dataset supports analyses of research trajectories for SED doctorate recipients and how they differ across subgroups (e.g., by degree field, demographics, etc.), which were previously infeasible using either source alone—offering new insights into participation in federally funded research.

Lessons learned from this project include the need for comprehensive DSAs, robust data safeguards, tailored tool selection, and actionable recommendations for future PPRL efforts. The successful implementation of PPRL with Anonlink in this context demonstrates its adaptability to constrained data environments and its potential for broader application across federal data linkage initiatives.

References

Brown, A. P., Randall, S. M., Boyd, J. H., and Ferrante, A. M. (2019). *Evaluation of approximate comparison methods on Bloom filters for probabilistic linkage.* International Journal of Population Data Science, 4(1), Article 16. https://doi.org/10.23889/ijpds.v4i1.1095

Christen, P., Hand, D.J, and Kirielle, N. (2023). *A review of the F-measure: its history, properties, criticism, and alternatives.* ACM Computing Surveys 56, no. 3 (2023): 1-24. https://dl.acm.org/doi/pdf/10.1145/3606367

CSIRO's Data61 (2017). *Anonlink: Private record linkage system*. GitHub. https://github.com/data61/anonlink

CSIRO's Data61 (2020). *Tutorial for CLI Tool anonlink-client.* https://anonlink-client.readthedocs.io/en/latest/tutorial/tutorial_cli.html

Datavant (2022). *Core Token Definitions and Rationale*. Unpublished internal document.

Datavant (2023-a). *Datavant Secures FedRAMP® Authorization, Trusted by Federal Agencies Through Privacy-Preserving Infrastructure*. Datavant. https://www.datavant.com/press-release/datavant-secures-fedramp-r-authorization-trusted-by-federal-agencies-through-privacy-preserving-infrastructure

Datavant (2023-b). *What is Privacy-Preserving Record Linkage (PPRL) and Why Does It Matter?* https://www.datavant.com/blog/privacy-preserving-record-linkage

U.S. Department of Health & Human Services (2025). *Summary of the HIPAA Privacy Rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

Karapiperis, D. and Verykios, V. (2015) *An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage*. IEEE Transactions on Knowledge & Data Engineering, vol. 27, no. 04, pp. 909-921. doi: 10.1109/TKDE.2014.2349916

Mirel L.B., Resnick D., Aram J., Cox C. (2022) *A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data.* Stat J IAOS. 2022 Jun 7;38(2):413-421. doi: 10.3233/sji-210891. PMID: 35910693; PMCID: PMC9335262.

National Institute of Standards and Technology (NIST) (2015). *Secure Hash Standard (SHS).* Federal Information Processing Standards Publication 180-4. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf

National Center for Science and Engineering Statistics (NCSES) (2022). *Matching SDR respondents to investigators of NSF awards.* NCSES Working Paper No. 22-211. U.S. National Science Foundation.

https://ncses.nsf.gov/pubs/ncses22211#use-cases_comparison-of-survey-and-administrative-data-on-nsf-support

NORC at the University of Chicago (2024). *PPRL Software Selections Recommendations Report.* PPRL2-23-N02, June 2024

NORC at the University of Chicago (2025-a). *Linkage Approach Report.* PPRL2-23-N02, August 2025

NORC at the University of Chicago (2025-b). *Statistical Analysis Report.* PPRL2-23-N02, September 2025

NORC at the University of Chicago (2025-c). *Data Enclave.* https://www.norc.org/services-solutions/data-enclave.html

Office of the Assistant Secretary for Planning and Evaluation (ASPE) (2024). *Childhood Obesity Data Initiative (CODI): Integrated data for patient-centered outcomes research project*. U.S. Department of Health & Human Services. https://aspe.hhs.gov/childhood-obesity-data-initiative-codi-integrated-data-patient-centered-outcomes-research-project

Office of Management and Budget (OMB) (2023). *Fundamental Responsibilities of Recognized Statistical Agencies and Units.* Notice of Proposed Rulemaking No. 2023-17664; 88 Fed. Reg. 56708–56744. Federal Register. https://www.federalregister.gov/documents/2023/08/18/2023-17664/fundamental-responsibilities-of-recognized-statistical-agencies-and-units

Resnick, D. and Adell Raventós, N. (2024) Efficiency and Privacy in Record Linkage: Evaluating a Novel Blocking Technique Implemented on Cryptographic Long-term Keys. International Population Data Linkage Conference. September, 2024.

Rosulek, Mike, et al. (2021) *The Joy of Cryptography.* Creative Commons. BY-NC-SA.

Schnell, R., Bachteler, T., and Reiher, J. (2011). *A Novel Error-Tolerant Anonymous Linking Code*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3549247